(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－93286

(43)公開日　平成７年(1995)４月７日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/00					
Ｇ０１Ｎ　33/68					
Ｇ０６Ｆ　17/30					
		8724－5Ｌ	Ｇ０６Ｆ	15/ 20	Ｄ
		9194－5Ｌ		15/ 40	530　Ｓ

審査請求　有　請求項の数７　ＯＬ　（全９頁）

(21)出願番号　特願平5－233822

(22)出願日　平成５年(1993)９月20日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　馬見塚　拓
東京都港区芝五丁目７番１号　日本電気株式会社内

(72)発明者　安倍　直樹
東京都港区芝五丁目７番１号　日本電気株式会社内

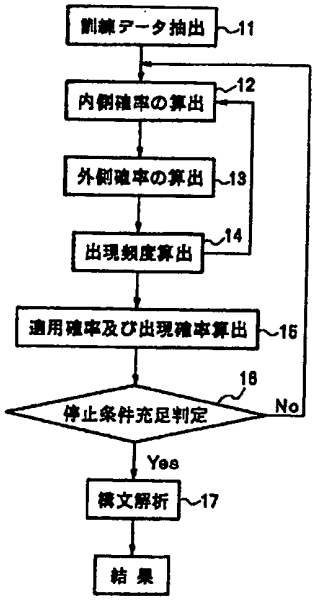
(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】　タンパク質立体構造予測方法

(57)【要約】

【目的】　構造未知のタンパク質のアミノ酸配列情報から、その立体構造内に含まれる長距離相互作用に基づく立体構造を高精度で予測する。

【構成】　ステップ１１で、構造既知及び未知のタンパク質アミノ酸配列から長距離相互作用を有する立体構造の訓練データを抽出し、ステップ１２で、木構造の生成確率における内側確率を算出し、ステップ１３で、木構造の生成確率における外側確率を算出し、ステップ１４でこれら内側確率及び外側確率から書き換え規則の出現頻度、及び各アミノ酸の出現頻度を算出し、ステップ１５で、書き換え規則の適用確率及び規則の末端ノードの各アミノ酸の出現確率を算出し、ステップ１６で、訓練データの反復学習の停止条件の充足判定を行い、ステップ１７で、学習により得られた書き換え規則を使用して、構文解析により立体構造未知のデータに対して、長距離相互作用を有する立体構造部位の予測を行う。



1

【特許請求の範囲】

【請求項1】　タンパク質のアミノ酸配列からタンパク質の構造予測を行うための訓練データを抽出するステップと、訓練データからタンパク質の部分的な立体構造に相当する書き換え規則を学習するステップと、学習された書き換え規則を用いて、テストアミノ配列データに対し、立体構造部分の予測を行うステップとからなることを特徴とするタンパク質立体構造予測方法。

【請求項2】　前記訓練データを抽出するステップは、立体構造既知のタンパク質に対し、同じファミリーに属するタンパク質、もしくは一次構造上、一定値以上の相同性を有するタンパク質のアミノ酸配列を、アミノ酸配列データベースから抽出することを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項3】　前記書き換え規則を学習するステップにおける書き換え規則が確率的な規則であり、該ステップは、確率的脈自由文法の学習に使用されるインサイド・アウトサイドアルゴリズムと呼ばれる学習方法の木文法への拡張であることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項4】　前記書き換え規則を学習するステップにおける書き換え規則が確率的な規則であり、末端ノードに20種類のアミノ酸もしくは、それらのカテゴリーが割り当てられ、それらの出現確率付き確率的規則であり、該ステップは、確率的文脈自由文法の学習に使用されるインサイド・アウトサイドアルゴリズムと呼ばれる学習方法の木文法への拡張であることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項5】　前記立体構造部分の予測を行うステップは、文脈自由文法の構文解析に使用されるCKYアルゴリズムと呼ばれる構文解析方法の拡張であることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項6】　前記立体構造部分の予測を行うステップにおける書き換え規則が確率的な規則であり、該ステップは、確率的文脈自由文法の構文解析に使用されるCKYアルゴリズムと呼ばれる構文解析方法の木文法への拡張であることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【請求項7】前記立体構造部分の予測を行うステップにおける書き換え規則が確率的な規則であり、末端ノードに20種類のアミノ酸もしくは、そのカテゴリーが割り当てられ、出現確率付き確率的規則であり、該ステップは、確率的文脈自由文法の構文解析に使用されるCKYアルゴリズムと呼ばれる構文解析方法の木文法への拡張であることを特徴とする請求項1記載のタンパク質立体構造予測方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、立体構造未知のタンパク質アミノ酸配列から、タンパク質の立体構造を予測する方法に関する。

2

【0002】

【従来の技術】タンパク質の立体構造を予測する方法としては、タンパク質全体の立体構造ではなく、その部分的な立体構造である二次構造を予測する方法が一般的である。従来、タンパク質二次構造予測問題は、タンパク質の一次構造の各残基（以下、予測対象となる残基を中心残基と呼ぶ）が、αヘリックス、βシート、それ以外という3種類の二次構造のいずれに対応するかを予測する問題として扱われてきた。従来技術によるタンパク質の二次構造を予測する方法として、例えば、1974年発行の米国の雑誌「バイオケミストリー」（Biochemistry）の第23巻222－245頁記載のチョウ（Chou）とファスマン（Fasman）による論文「プレディクション オブ プロテイン コンホメーション」（Prediction of protein conformation）（以下、CF法と略す）、1978年発行の米国の雑誌「ジャーナル オブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第120巻97－120頁掲載のガルニエ（Garnier）らによる論文「アナリシス オブ ザ アキュレシー アンド インプリケーションズ オブ シンプル メソード フォー プレディクティング ザ セコンダリー ストラクチャー オブ グロブラー プロテインズ」（Analysisof the accuracy and implications ofsimple method for predicting the secondary structure of globular proteins）（以下、GOR法と略す）、1987年発行の米国の雑誌「ジャーナルオブ モレキュラ バイオロジー」（Journal of Molecular Biology）の第198巻425－443頁掲載のギブラト（Gibrat）らによる論文「ファザー デベロプメンツ オブ プロテイン セコンダリー ストラクチャ プレディクション ユージング インホメーション セオリー：ニュー パラメータズ アンド コンシダレーション オブ レジデューペアズ」（Further developments of proteinsecondary structure prediction using information theory:New paramentersand consideration of residue pairs）（以下、GGR法と略す）、1988年発行の米国の雑誌「ジャーナル オブモレキュラ バイオロジー」（Journal of Molecular Biology）の第202巻865－884頁記載のキャン（Qian）らによる論文「プレデイクティング ザ セコンダリー ストラクチャー オブ グロブラー プロテインズ ユ

ージング　ニューラル　ネットワーク　モデルズ」（Predicting the secondary structure of globular proteins using neural network models）（以下、ＱＳ法と略す）などがある。

【０００３】ＣＦ法は、タンパク質構造のデータベースから各二次構造におけるアミノ酸の統計的な出源頻度を求め、この頻度表を使用し、経験的な規則に基づく予測を行っている。また、ＧＯＲ法は、中心残基の二次構造に対して、その残基から数残基離れた残基により独立にもたらされる情報量の和を計算し、その相対値から予測を行い、ＧＧＲ法は、中心残基の二次構造に対して、その残基及びその残基から数残基から数残基離れた残基によりもたらされる情報量の和から予測を行っている。さらにＱＳ法は、３層のフィードフォワード型のネットワークを使用し、中心残基の前後８残基を含む配列を入力とし、二次構造に対する中心残基及び周辺残基からの寄与をニューラルネットワークを用いて抽出することにより予測を行っている。

【０００４】

【発明が解決しよとする課題】タンパク質の立体構造の中には、その部分構造においても、数残基から数十残基以上も離れたアミノ酸残基同士の相互作用（以下、遠距離相互作用と呼ぶ）により構成されているものが数多くある。例えば、主要な二次構造の一つであるβシートも、その一つである。しかし、従来、タンパク質立体構造予測、例えば、βシートの予測を含む二次構造予測手法などにおいて、そのような遠距離相互作用を保持している残基同士の間に存在するアミノ酸残基を無視し、遠距離相互作用を自動的に抽出する方法や、さらに、その遠距離相互作用を規則として未知データに対する予測を行う方法は皆無であり、そういった手法は確立されていなかった。

【０００５】

【課題を解決するための手段】本発明のタンパク質立体構造予測方法は、タンパク質のアミノ酸配列からタンパク質の構造予測を行うための訓練データを抽出するステップと、訓練データからタンパク質の部分的な立体構造に相当する書き換え規則を学習するステップと、学習された書き換え規則を用いて、テストアミノ酸配列データに対し、立体構造部分の予測を行うステップとからなることを特徴とする。

【０００６】又、前記訓練データを抽出するステップは、立体構造既知のタンパク質に対し、同じファミリーに属するタンパク質、もしくは一次構造上、一定値以上の相同性を有するタンパク質のアミノ酸配列を、アミノ酸配列データベースから抽出することを特徴とする。

【０００７】又、前記書き換え規則を学習するステップにおける書き換え規則が確率的な規則であり、該ステップは、確率的文脈自由文法の学習に使用されるインサイド・アウトサイドアルゴリズムと呼ばれる学習方法の木文法への拡張であることを特徴とする。

【０００８】又、前記書き換え規則を学習するステップにおける書き換え規則が確率的な規則であり、末端ノードに２０種類のアミノ酸もしくは、それらのカテゴリーが割り当てられ、それらの出現確率付き確率的規則であり、該ステップは、確率的文脈自由文法の学習に使用されるインサイド・アウトサイドアルゴリズムと呼ばれる学習方法の木文法への拡張であることを特徴とする。

【０００９】又、前記立体構造部分の予測を行うステップは、文脈自由文法の構文解析に使用されるＣＫＹアルゴリズムと呼ばれる構文解析方法の拡張であることを特徴とする。

【００１０】又、前記立体構造部分の予測を行うステップにおける書き換え規則が確率的な規則であり、該ステップは、確率的文脈自由文法の構文解析に使用されるＣＫＹアルゴリズムと呼ばれる構文解析方法の木文法への拡張であることを特徴とする。

【００１１】又、前記立体構造部分の予測を行うステップにおける書き換え規則が確率的な規則であり、末端ノードに２０種類のアミノ酸もしくは、そのカテゴリーが割り当てられ、出現確率付き確率的規則であり、該ステップは、確率的文脈自由文法の構文解析に使用されるＣＫＹアルゴリズムと呼ばれる構文解析方法の木文法への拡張であることを特徴とする。

【００１２】

【実施例】次に、本発明について図面を参照して詳細に説明する。本実施例では、対象とする遠距離相互作用からなるタンパク質立体構造としてβシート領域を扱うものとする。

【００１３】図１は、本発明において遠距離相互作用を保持した立体構造の規則に相当する書き換え規則の例である。一般に、書き換え規則は、非終端記号及び終端記号がラベル付けされたノードを有する木であり、固有の適用確率を持つ。書き換え規則の書き換え動作は、あらかじめ設定した初期木に対して、書き換え規則を有限回適用し、終端記号のみからなる木構造を構成することで達成される。

【００１４】ここで、木構造とは、単一ノード、もしくは複数の木構造を左右に順序づけられた子供として持つノードからなる構造として再帰的に定義される。また、木構造のルートノードとは、そのノードを子供とするノードが存在しない木構造の中の唯一のノードを指す。さらに、ノードのランクとは、ノードの下の子供の数を指す。例えば、図１はランク１の書き換え規則である。

【００１５】具体的に、生成確率P_Aを持つ木構造Ａに含まれるランクｒの非終端記号によりラベル付けされたノードＴを、適用確率P_Bを有する同ランクの木構造Ｂによって書き換えるということは、Ａ中のＴをＢによって置き換え、Ｔのｒ個の子供を各々Ｂ中のｒ個の空白ノ

ードの位置に、左から順番に連結し、新たに導出された木構造とし、さらに、P_A とP_B との積をとることにより、新しい生成確率を計算することを指す。

【００１６】ここで、終端記号は、２０種類のアミノ酸、もしくは、アミノ酸をその化学的性質に基づいて分類したいくつかのグループに対応し、導出された木構造の末端に現れるそれらアミノ酸もしくはグループの集合はアミノ酸配列に相当する。与えらえたアミノ酸配列に対し、適当な書き換え規則を使用し、そのアミノ酸配列が木構造の末端に出現するように書き換え動作を行うことにより、その配列上のどの部分が各書き換え規則により生成されたかを調べることを「構文解析（ｐａｒｓｉｎｇ）」という。

【００１７】また、タンパク質アミノ酸配列の各残基位置においては、１つのアミノ酸種類に固定されず、他のアミノ酸に置き換えられても、その立体構造及び機能を保持していることが多い。そのため、一般的な書き換え規則をそのまま使用した場合、全体では類似していながら、末端文字のみが異なる木構造が数多く出現する。そこで、あらかじめ、書き換え規則を、その末端にラベル付けされている一つの終端記号を２０種類のアミノ酸と各々に付随した２０の出現確率とにそれぞれ置き換え、出現確率を加味した書き換え規則（以下、出現確率付き書き換え規則）にしておいた方が、書き換え規則の数が減り、計算上都合が良く、また、規則も見やすくなる。この場合、生産確率P_A を持つ木構造Ａに含まれるランクｒの非終端記号によりラベル付けされたノードＴを、適用確率P_B を有する同ランクの出現確率付き確率規則Ｂによって書き換えた場合、P_A とP_B とＢの各末端ノードの対応するアミノ酸に付随した出現確率の積をとることにより、新しい生成確率を計算する。出現確率付き書き換え規則の例を図２に示す。

【００１８】以下、ランク１の出現確率付き書き換え規則において、１種類の非終端記号をルートノードに使用する場合について、βシート領域の規則の学習方法、及び規則の適用による予測方法について具体的に説明する。

【００１９】図３は、本発明のタンパク質立体構造予測方法の実施例を説明するフローチャートである。

【００２０】ステップ１１では、まず、βシート領域既知のタンパク質をタンパク質立体構造データベースから抽出する。さらに、該タンパク質に対し、同じタンパク質あるいは同じファミリーに属するタンパク質、もしくは、一次構造上、一定の割合で相同性を保持しているタンパク質のアミノ酸配列をタンパク質アミノ酸配列データベースから抽出し、学習に使用するデータとする。配列データベースから抽出したタンパク質においては、βシート領域が未知でも構わないとする。

【００２１】例えば、イミノグロブリンというタンパク質において、ヒトのそのタンパク質のアミノ酸配列上のどこにβシート領域が存在するかは物理化学的な実験から明らかになっている。このヒト・イミノグロブリンに対して、ヒト以外の種類、例えば、チンパンジー、イヌ等のイミノグロブリンのアミノ酸配列、あるいは、アライメントにより一定の割合以上の相同性を有するアミノ酸配列を学習データとする。

【００２２】ステップ１２、１３、１４、及びステップ１５は、ステップ１１で得られた学習データを使用し、あらかじめ設定した構造を有する書き換え規則の適用確率及び書き換え規則の末端ノードにおける各アミノ酸の出現確率を学習するステップである。

【００２３】ステップ１２は内側確率の算出を行うステップ、ステップ１３は外側確率の算出を行うステップ、ステップ１４は内側確率、外側確率から、書き換え規則の適用確率及びその末端ノードにおける各アミノ酸の出現確率の計算に必要な、書き換え規則の出現頻度及び末端ノードの各アミノ酸の出現頻度を計算するステップである。ステップ１５は、算出された出現頻度を使用し、書き換え規則の適用確率及びその末端ノードの各アミノ酸の出現確率を計算する。

【００２４】ステップ１２の内側確率の算出過程を、図４に示したフローチャートを用いて説明する。内側確率の算出には、４次元のテーブルＩｎ［ｉ，ｊ，ｋ，ｌ］を用意する。テーブルＩｎの次元は書き換え規則のランクに依存し、書き換え規則のランクｒに対し、２（ｒ＋１）次元のテーブルが必要である。テーブルＩｎ［ｉ，ｊ，ｋ，ｌ］においては、有限回の書き換え規則の適用により生成された木構造の末端に出現するアミノ酸配列が、与えられた訓練データのｉ番目からｊ番目までの残基位置、ｋ番目から１番目までの残基位置に相当しており、それらすべての木構造が生成された生成確率の和を示す。訓練データの各配列が与えられる前に、それぞれＩｎ［ｉ，ｊ，ｋ，ｌ］＝０．０に設定しておく。

【００２５】訓練データの配列が与えられた場合に、ステップ２１において、テーブルＩｎ［ｉ，ｊ，ｋ，ｌ］を初期化する。初期化は、訓練データに対し、書き換え規則を１回だけ適用し、得られた木構造の末端ノードと訓練データの部分配列とを対応させることを指す。この動作により、生成された木構造の末端に出現するアミノ酸配列が、与えられたアミノ酸配列の部分配列に相当し、対応するテーブルＩｎ［ｉ，ｊ，ｋ，ｌ］に生成確率を格納する。もし、他の書き換え規則の適用により生成された部分配列に相当する。アミノ酸配列の位置が、同様にｉ，ｊ，ｋ，ｌであれば、この書き換え規則の生成確率を、テーブルＩｎ［ｉ，ｊ，ｋ，ｌ］に加算する。

【００２６】例えば、長さが４以上のアミノ酸配列に、図２に示す書き換え規則を適用した場合に生成された部分配列に相当するアミノ酸配列の位置としては、例えば、ｉ＝１，ｊ＝２ｋ＝３，ｌ＝４が考えられる。この

時、書き換え規則の各末端ノードに出現するアミノ酸の出現確率及び書き換え規則の適用確率のすべての積をとったものがテーブルＩｎ［1，2，3，4］の値となる。

【００２７】以上のように、与えられた訓練データにおいて、書き換え規則の適用により生成された部分配列に相応する、取り得るすべてのアミノ酸配列の位置のテーブルの初期化を行う。

【００２８】次に、与えられた訓練データに対して取り得るすべてのテーブルＩｎに格納する値の計算を行う。

【００２９】まず、ステップ２２において、訓練データのアミノ酸配列の長さがＮであれば、ｉ＝Ｎ，ｊ＝ｉ，ｋ＝Ｎ，ｌ＝ｋと設定する。さらに、ステップ２３において、ｉ，ｊ，ｋ，ｌの値を動かしながら、各ｉ，ｊ，ｋ，ｌの値において、ステップ２４の動作を行う。

【００３０】ステップ２３におけるｉ，ｊ，ｋ，ｌの動作を説明する。ｉを１になる（ｉ＝１）まで１ずつ減らし（ｉ＝ｉ－１）、各ｉにおいてｊをＮになる（ｊ＝Ｎ）まで１ずつ増やし（ｊ＝ｊ＋１）、各ｊにおいてｋをｊになる（ｋ＝ｊ）まで１ずつ減らし（ｋ＝ｋ－１）、各ｋにおいてｌがＮになる（ｌ＝Ｎ）まで、ｌの値を１ずつ増や（ｌ＝ｌ＋１）していく。

【００３１】ステップ２４においては、各ｉ，ｊ，ｋ，ｌで、すべての書き換え規則の末端ノード数を調べ、テーブルＩｎ［ｉ，ｊ，ｋ，ｌ］への書き換え規則の適用により得られた木構造に相当するテーブルＩｎの値が０．０ではない場合にのみ、生成確率の計算を行ないテーブルＩｎ［ｉ，ｊ，ｋ，ｌ］に格納する。もし、複数の書き換え規則により、テーブルＩｎ［ｉ，ｊ，ｋ，ｌ］の生成確率が算出されれば、それらの和をテーブルＩｎ［ｉ，ｊ，ｋ，ｌ］に格納する。

【００３２】例えば、ｉ＝３，ｊ＝７，ｋ＝８，ｌ＝１０であり、ある書き換え規則の末端ノード数がそれぞれ１，２，０，１であれば、テーブルＩｎ［４，５，８，９］に０．０でない値が格納されている時、訓練データのアミノ酸配列のそれぞれ３，６，７，１０番目のアミノ酸に対応する末端ノードでの各アミノ酸の出現確率と書き換え規則の適用確率とテーブルＩｎ［４，５，８，９］の積を計算し、その値を生成確率としてテーブルＩｎ［３，７，８，１０］に加算する。

【００３３】ステップ２５では、ステップ２４の動作の終了の判断を行う。ステップ２３によりｉ，ｊ，ｋ，ｌの値がｉ＝１，ｊ＝Ｎ，ｋ＝Ｎ，ｌ＝Ｎとなった場合、ステップ２４の動作後、内側確率の計算を終了する。

【００３４】以上の動作により、取り得るすべてのＩｎ［ｉ，ｊ，ｋ，ｌ］が計算でき、ステップ１２の内側確率の算出を終了する。

【００３５】次に、ステップ１３での外側確率の算出過程を図５に示したフローチャートを用いて説明する。

【００３６】内側確率の算出と同様に、外側確率の算出においても、４次元のテーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］を使用する。テーブルＯｕｔの次元は、Ｉｎと同様に書き換え規則のランクに依存する。テーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］は、有限回の書き換え規則の適用により生成された木構造の末端ノードに出現するアミノ酸配列が、与えらえたＮ残基からなる訓練データの１番目からｉ番目までの残基位置、ｊ番目からｋ番目までの残基位置、ｌ番目からＮ番目までの残基位置に相当しており、生成された木構造の生成確率の和を示す。各テーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］は、訓練データの各配列が与えられる前に、０．０に設定しておく。

【００３７】訓練データの配列が与えられた場合に、ステップ３１において、テーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］を初期化する。初期化は、初期木に対して書き換え規則を１回のみ適用することを指す。この動作により、生成された木構造の末端に出現するアミノ散配列が、与えられたアミノ酸配列の部分配列に相当し、対応するテーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］に生成確率を格納する。もし、他の書き換え規則の適用により生成された部分配列の訓練データ上の位置が、同様にｉ，ｊ，ｋ，ｌであれば、この書き換え規則各末端ノードに出現するアミノ酸の出現確率及び書き換え規則の適用確率のすべての積を、テーブルＯｕｔ［ｉ，ｊ，ｋ，ｌ］に加算する。

【００３８】例えば、長さが１０の訓練データに、図２に示す書き換え規則の適用により生成された部分配列に相当するアミノ酸配列の位置は、ｉ＝１，ｌ＝１０でなければならず、さらにｊ，ｋに関しては、ｋ＝ｊ＋１を満たす７通りが考えられる。例えば、ｉ＝１，ｊ＝５，ｋ＝６，ｌ＝１０であれば、各位置のアミノ酸が書き換え規則の各末端ノードに出現する出現確率及び書き換え規則の適用確率のすべての積を算出したものをテーブルＯｕｔ［１，５，６，１０］に加算する。

【００３９】以上のように、与えられた訓練データにおいて、書き換え規則の適用により生成された部分配列に相当する、取り得るすべてのアミノ酸配列の位置に対応するテーブルＯｕｔの初期化を行う。

【００４０】次に、与えられた訓練データに対して取り得るすべてのテーブルＯｕｔに格納する値の計算を行う。

【００４１】まず、ステップ３２において、訓練データのアミノ酸配列の長さがＮであれば、ｉ＝１，ｊ＝Ｎ，ｋ＝ｊ，ｌ＝Ｎと設定する。さらに、ステップ３３において、ｉ，ｊ，ｋ，ｌの値を動かしながら、各ｉ，ｊ，ｋ，ｌにおいて、ステップ３４の動作を行う。

【００４２】ステップ３３におけるｉ，ｊ，ｋ，ｌの動作を説明する。ｉをＮになる（ｉ＝Ｎ）まで１ずつ増やし（ｉ＝Ｎ）、各ｉにおいてｊをｉになる（ｊ＝ｉ）まで１ずつ減らし（ｊ＝ｊ－１）、各ｊにおいてｋをＮになる（ｋ＝Ｎ）まで１ずつ増やし（ｋ＝ｋ＋１）、各ｋ

において l がkになる（l＝k）まで、 l の値を1ずつ減ら（l＝l－1）していく。

【0043】ステップ34においては、各書き換え規則の末端ノード数を調べ、その書き換え規則を適用した場合に、生成された木構造がテーブルOut［i，j，k，l］に対応するような木構造が存在する場合にのみ、生成確率の計算を行ないテーブルOut［i，j，k，l］に格納する。もし複数の書き換え規則により、テーブルOut［i，j，k，l］の生成確率が算出されれば、それらの和をテーブルOut［i，j，k，l］に格納する。

【0044】例えば、i＝3，j＝4，k＝7，l＝9であり、書き換え規則の末端ノード数がそれぞれ1，2，0，1であれば、テーブルOut［2，6，7，10］に値が格納されている時、訓練データのアミノ酸配列のそれぞれ3，4，5，9番目のアミノ酸に対応する末端ノードの各アミノ酸の出現確率と書き換え規則の適用確率とOut［2，6，7，10］の積を計算し、その値を生成確率としてテーブルOut［3，4，7，9］に加算する。

【0045】ステップ35では、ステップ34の動作の終了の判断を行う。ステップ33により、i，j，k，lの値がi＝N，j＝i，k＝N，l＝kとなった場合、ステップ34の動作後、外側確率の計算を終了する。

【0046】以上の動作により、取り得るすべてのOut［i，j，k，l］が計算でき、ステップ13の外側確率の算出を終了する。

【0047】次に、ステップ14において、内側確率、外側確率から、書き換え規則の出現頻度、及びその末端ノードの各アミノ酸の出現頻度を計算する。

【0048】ステップ14の出現頻度の算出過程を図6に示したフローチャートを用いて説明する。まず、出現頻度を格納する4次元のテーブルPd［m，n，p，q］を用意する。テーブルPdの次元は、In及びOutとは異なり、書き換え規則のランクに依存しない。20種類のアミノ酸を1から20までのアミノ酸番号に置き換えた場合に、テーブルPd［m，n，p，q］の添え字m，n，p，qは、m番目の書き換え規則において、非終端記号のn番目のノード位置のp番目の末端ノードに出現するq番目のアミノ酸を示す。各テーブルPd［m，n，p，q］の値は、訓練データの最初の配列が与えられた時にのみあらかじめ0．0に設定しておく。

【0049】ステップ41においては、i＝1，j＝N，k＝j，l＝Nと設定する。さらに、ステップ42においてi，j，k，lの値を動かしながら各i，j，k，lにおいて、ステップ43の動作を行う。

【0050】ステップ42でのi，j，k，lの動作を説明する。iをNになる（i＝N）まで1ずつ増やし（i＝N）、各iにおいてjをiになる（j＝i）まで1ずつ減らし（j＝j－1）、各jにおいてkをNになる（k＝N）まで1ずつ増やし（k＝k＋1）、各kにおいて l がkになる（l＝k）まで、 l の値を1ずつ減ら（l＝l－1）していく。

【0051】ステップ43においては、各書き換え規則それぞれに対し、テーブルOut［i，j，k，l］に対応する木構造に書き換え規則を適用した場合に、対応する添え字を要素とするテーブルInが0．0ではない値を有して存在するかどうかをチェックする。存在していれば、そのテーブルInとOut［i，j，k，l］との間を埋める部分配列に対応する書き換え規則の各末端ノードの各アミノ酸の出現確率、書き換え規則の適応確率、テーブルOut［i，j，k，l］、Inの積を計算し、テーブルPdに加算する。

【0052】具体的に、i＝2，j＝6，k＝7，l＝12の時、図2の構造をした適用確率Pを有する書き換え規則を1番として適用する場合を考える。図2の書き換え規則においては、4個のノード位置に一つずつノードが存在するので、各々訓練データの3，5，8，11番目の残基位置のアミノ酸に対応する。また、この位置のアミノ酸はそれぞれ、アミノ酸番号により、2，15，18，7番であり、書き換え規則の対応する各ノードのアミノ酸の出現確率は、それぞれp1，p2，p3，p4であるとする。この時、Pd［1，1，1，2］，Pd［1，2，1，15］，Pd［1，3，1，18］，Pd［1，4，1，7］それぞれに、P×p1×p2×p3×p4×Out［2，6，7，12］×In［4，4，9，10］が加算される。

【0053】ここで、テーブルInの添え字は、テーブルOutの添え字と書き換え規則のノードの単純な差ではないことに注意する。

【0054】ステップ44では、ステップ43の動作の終了の判断を行う。ステップ42により、i，j，k，lの値がi＝N，j＝i，k＝N，l＝kとなった場合、ステップ43の動作後、テーブルPdの算出を終了する。

【0055】訓練データの各配列に対して、ステップ12、13、14を繰り返し、書き換え規則及び規則の各末端ノード位置での各アミノ酸の出現頻度を算出する。

【0056】ステップ15では、算出を行ったテーブルPdから、書き換え規則の適用確率、及び各末端ノードの各アミノ酸の出現確率を計算する。

【0057】M番目の書き換え規則の適用確率は、

【0058】

【数1】

$$\frac{\sum_n \sum_p \sum_q Pd[M,n,p,q]}{\sum_m \sum_n \sum_p \sum_q Pd[m,n,p,q]}$$

【0059】により計算される。また、M番目の書き換

え規則のN番目のノード位置のP番目のノードにQ番のアミノ酸が出現する出現確率は、

【0060】

【数2】

$$\frac{Pd[M,N,P,Q]}{\sum_q Pd[M,N,P,q]}$$

【0061】により計算される。

【0062】以上により、訓練データセットに対して、あらかじめ設定した書き換え規則の適用確率及び各書き換え規則の末端ノード位置にアミノ酸が出現する出現確率が算出される。

【0063】ステップ12、13、14、15までをあらかじめ設定した回数、もしくは、あらかじめ設定した停止条件を満たすまで繰り返す。ステップ16は、停止条件が満たされているかどうかをチェックする。例えば、停止条件としては、「いずれの書き換え規則の適用確率やアミノ酸の出現確率の値も変化が0．01未満である」などが採用できる。

【0064】ステップ17では、与えられたテストアミノ酸配列に対し、書き換え規則の適用確率及び規則の各末端ノードのアミノ酸の出現確率から、構文解析により、どの書き換え規則の末端ノードがテスト配列の部分領域に対応するかを調べる。すなわち、構文解析により、遠距離相互作用を保持した書き換え規則がテストアミノ酸配列のどの部分に現れたかを検出し、テストアミノ酸配列内の遠距離相互作用を保持している部分を予測する。

【0065】ステップ17の構文解析方式を、図7に示したフローチャートを用いて説明する。まず、4次元のテーブルPar［i，j，k，l］を用意する。テーブルParの次元は書き換え規則のランクに依存し、書き換え規則のランクrに対し、2（r＋1）次元のテーブルを用意する。テーブルPar［i，j，k，l］は、書き換え規則の適用により生成された木構造の末端ノードに出現するアミノ酸配列が、与えられたN残基からなるテストデータの1番目からi番目までの残基位置、j番目からk番目までの残基位置、1番目からN番目までの残基位置に相当しており、生成された木構造の生成確率の最大値を示す。テストデータが与えられる前に、各テーブルPar［i，j，k，l］＝0．0と設定しておく。

【0066】テストアミノ酸配列が与えられた場合に、ステップ51において、テーブルPar［i，j，k，l］を初期化する。初期化は、初期木に対して書き換え規則を1回のみ適用することを指す。この動作により、生成された木構造の末端ノードのアミノ酸列が、与えられたアミノ酸配列の部分配列に相当し、対応するテーブルPar［i，j，k，l］に生成確率を格納する。もし、他の書き換え規則の適用により生成された部分配列のテストデータ上の位置が、同様にi，j，k，lであれば、この書き換え規則の各末端ノードに出現するアミノ酸の出現確率及び書き換え規則の適用確率の積を、Par［i，j，k，l］と比較し、大きい方をPar［i，j，k，l］の値とする。

【0067】次に、テストデータに対して、取り得るすべてのテーブルOut［i，j，k，l］に格納する値の計算を行う。

【0068】まず、ステップ52において、訓練データのアミノ酸配列の長さがNであれば、i＝1，j＝N，k＝j，l＝Nと設定する。さらに、ステップ53において、i，j，k，lの値を動かしながら、各i，j，k，lにおいて、ステップ54の動作を行う。

【0069】ステップ53におけるi，j，k，lの動作を説明する。iをNになる（i＝N）まで1ずつ増やし（i＝N）、各iにおいてjをiになる（j＝i）まで1ずつ減らし（j＝j－1）、各jにおいてkをNになる（k＝N）まで1ずつ増やし（k＝k＋1）、各kにおいてlがkになる（l＝k）まで、lの値を1ずつ減ら（l＝l－1）していく。

【0070】ステップ54においては、各書き換え規則の末端ノード数を調べ、その書き換え規則を適用した場合に、生成された木構造がテーブルPar［i，j，k，l］に対応するような木構造が存在する場合にのみ、生成確率の計算を行ないテーブルPar［i，j，k，l］に格納する。もし、複数の書き換え規則により、テーブルPar［i，j，k，l］の値が算出されれば、それらの中で最大の値をテーブルPar［i，j，k，l］に格納する。

【0071】ステップ55では、ステップ54の動作の終了の判断を行う。ステップ33により、i，j，k，lの値がi＝N，j＝i，k＝N，l＝kとなった場合、ステップ54の動作後、Parの計算を終了する。

【0072】ステップ56では、得られた木構造の末端ノードからなるアミノ酸配列が、与えられたテストデータに対応しているテーブルParの中で、最大のParを選出する。

【0073】ステップ57では、最大のParにおいて、それを算出するために、どの書き換え規則が使われてきたかをチェックする。このチェックは、ステップ54におけるParの算出において、軌跡を記憶しておくことによっても達成される。チェックの結果、例えば、テスト配列に対し、βシート領域に相当する書き換え規則が適用されていれば、適応により生成された部分配列は、βシート領域とアミノ酸配列レベルで非常に近い性質を保持しており、βシート領域である可能性が高いと予測する。

【0074】

【発明の効果】立体構造の既知のタンパク質のアミノ酸配列情報から、立体構造未知のタンパク質の遠距離相互

13

作用に由来する立体構造を従来技術に対して高い精度で予測可能である。すなわち、本手法により、遠距離に位置するアミノ酸残基同士の相互作用を、中間領域を介せず抽出可能であり、既存手法による局所領域からの予測では誤って予測されていたような領域を相互作用の有無という観点からより実際的に予測することが可能になった。また、本手法はβシートなどの部分的な立体構造として著名な部分のみならず、一次構造上離れた残基同士の相互作用により構成されている活性部位などの機能部位の特徴配列を抽出し、規則として予測に使用することが可能である。

【図面の簡単な説明】

【図1】本発明で使用する書き換え規則の一例を示す摸式図

【図2】本発明で使用する書き換え規則の一例を示す摸式図

【図3】本発明のタンパク質立体構造予測の一実施例を示すフローチャート

【図4】本発明の学習方式の一部である内側確率の算出方法の一実施例を示すフローチャート

【図5】本発明の学習方式の一部である外側確率の算出方法の一実施例を示すフローチャート

【図6】本発明の学習方式の一部である書き換え規則の出現頻度、及び規則の末端ノードの各アミノ酸の出現頻度の算出方法の一実施例を示すフローチャート

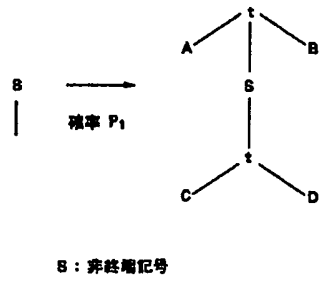
【図7】本発明の構文解析方式の一実施例を示すフローチャート

【符号の説明】

14

11　訓練データ抽出
12　内側確率の算出
13　外側確率の算出
14　出現頻度算出
15　適用確率及び出現確率算出
16　停止条件充足判定
17　構文解析
21　内側確率の初期化
22　初期添え字の設定
23　添え字の更新
24　内側確率の算出
25　停止条件充足判定
31　外側確率の初期化
32　初期添え字の設定
33　添え字の更新
34　外側確率の算出
35　停止条件充足判定
41　初期添え字の設定
42　添え字の更新
43　出現頻度の算出
44　停止条件充足判定
51　最大生成確率の初期化
52　初期添え字の設定
53　添え字の更新
54　最大生成確率の算出
55　停止条件充足判定
56　最大生成確率の算出
57　最大生成確率時の軌跡の検出

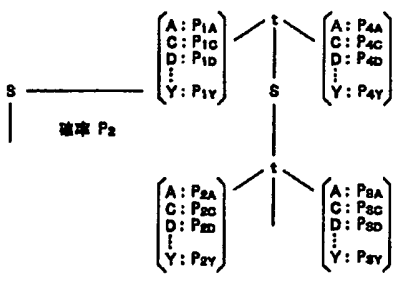
【図1】



S：非終端記号
A,B,C,D：終端記号（アミノ酸）
t：ノード

書きかえ規則の例

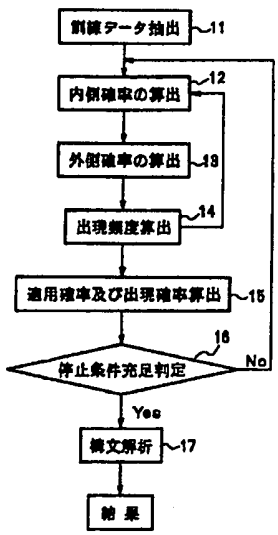
【図2】



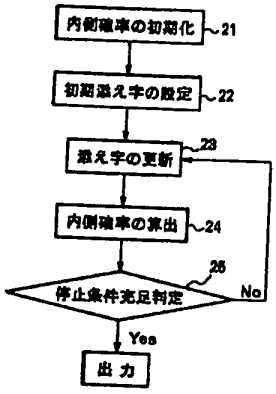
S：非終端記号
A,C,D,…Y：終端記号（アミノ酸）
t：ノード

出現確率つき書きかえ規則の例

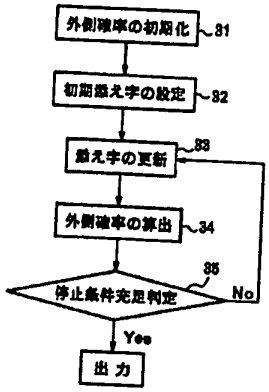
【図3】



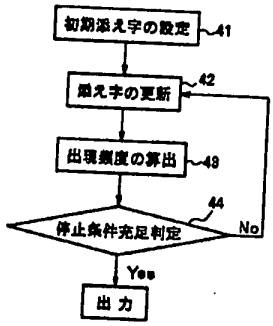
【図4】


内側確率の初期化
21
初期添え字の設定
22
添え字の更新
23
内側確率の算出
24
停止条件充足判定
25
No
Yes
出力

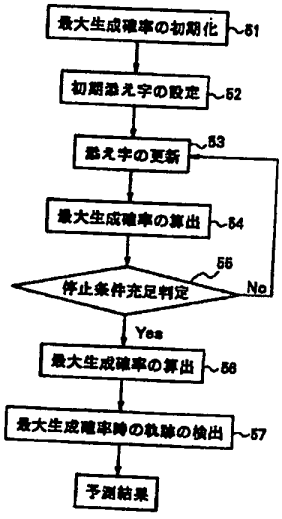
【図5】


外側確率の初期化
31
初期添え字の設定
32
添え字の更新
33
外側確率の算出
34
停止条件充足判定
35
No
Yes
出力

【図6】


初期添え字の設定
41
添え字の更新
42
出現頻度の算出
43
停止条件充足判定
44
No
Yes
出力

【図7】


最大生成確率の初期化
51
初期添え字の設定
52
添え字の更新
53
最大生成確率の算出
54
停止条件充足判定
55
No
Yes
最大生成確率の算出
56
最大生成確率時の軌跡の検出
57
予測結果